トランスクリプトーム解析の一手段としての次世代シーケンサーデータ解析
～実演を中心に～

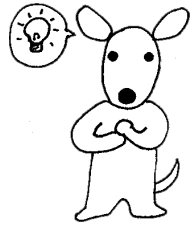
東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
門田 幸二（かどた　こうじ）
http://www.iu.a.u-tokyo.ac.jp/~kadota/
kadota@iu.a.u-tokyo.ac.jp

門田の部分は

- バイオインフォマティクス人材育成講座
  - 7月：**マイクロアレイ**受託実験や遺伝子発現データ解析
  - 8月：**次世代シーケンサー**見学
  - 8月：**統計解析言語R**や遺伝子発現データの**クラスタリング**

  } このあたりの補講

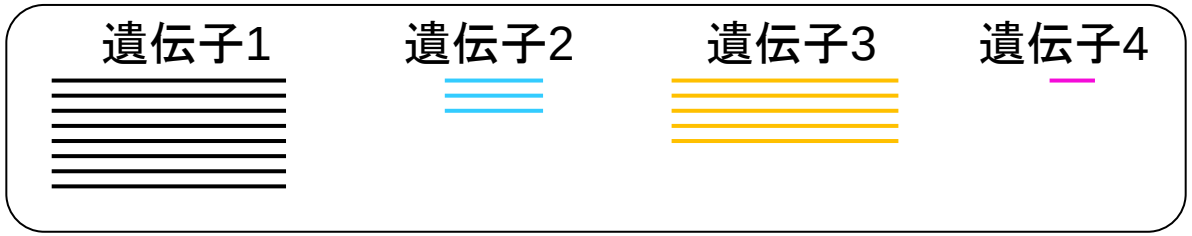
- 第一部のねらい
  - クラスタリング（やエントロピー）などのバイオインフォマティクスの**基本的なスキルを身につけるだけで様々な局面に応用可能である**という二つの事例を紹介するとともに、これらの具体的な計算例を示すことで数式アレルギー緩和に貢献



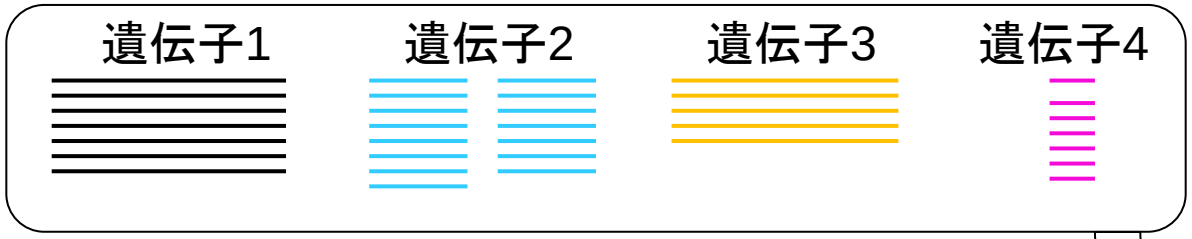
- 第二部のねらい
  - 次世代シーケンサーのデータだって、**統計解析言語Rでお手軽に解析**できる
  - **遺伝子発現行列**にしたら後は同じ
  - **高度なプログラミング能力がなくても**バイオインフォマティクスの世界で生存可能
  - 必要な情報はインターネットのみで十分（＠沖縄）

トランスクリプトーム情報を得る手段

- 光刺激前（T1）の目のトランスクリプトーム



- 光刺激後（T2）の目のトランスクリプトーム



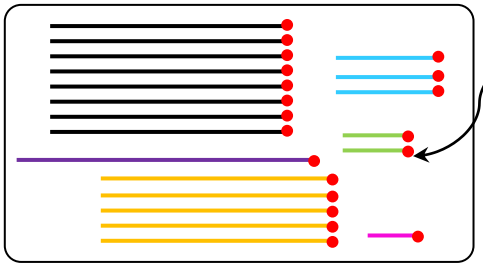
これがいわゆる
「遺伝子発現行列」

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

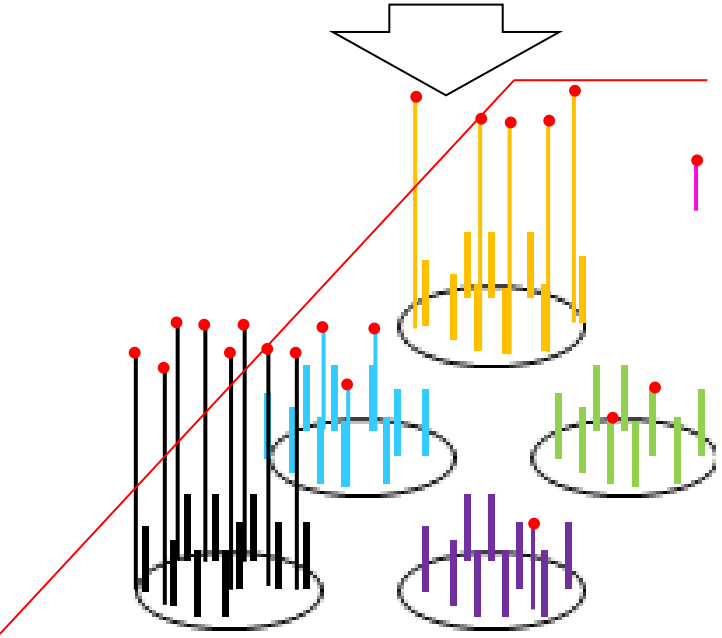
- マイクロアレイ
- 電気泳動に基づく方法
- 配列決定に基づく方法

マイクロアレイデータ → 遺伝子発現行列

- 光刺激前（T1）の目のトランスクリプトーム

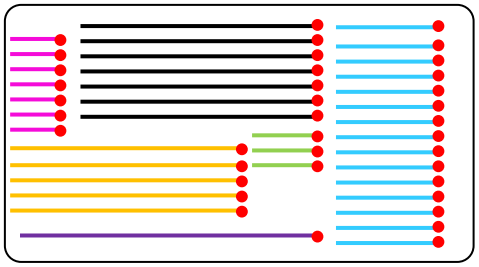


ハイブリダイゼーション
（二本鎖形成）



専用の検出器で各
遺伝子に対応する
領域の蛍光シグナ
ル強度を測定

光刺激後（T2）の目の
トランスクリプトーム

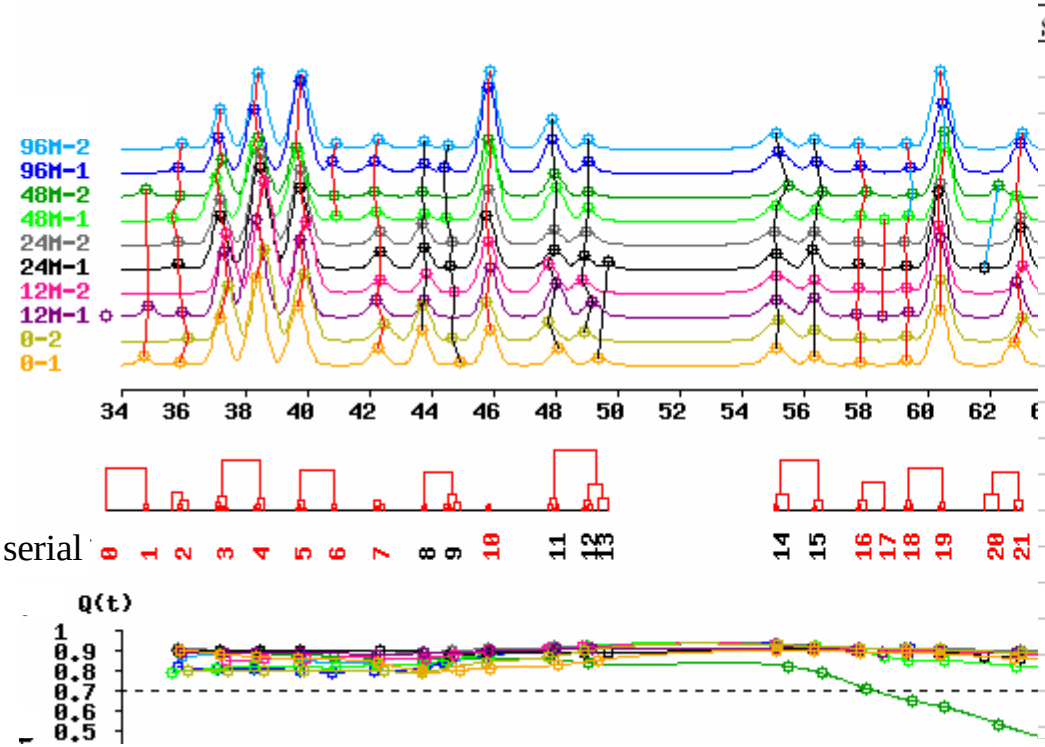


ハイブリダイゼーション
と
シグナル検出

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	?	?
遺伝子5	…	…
…	…	…

電気泳動データ → 遺伝子発現行列

- ピークのアラインメントがとれている　＝　遺伝子発現行列を作れている

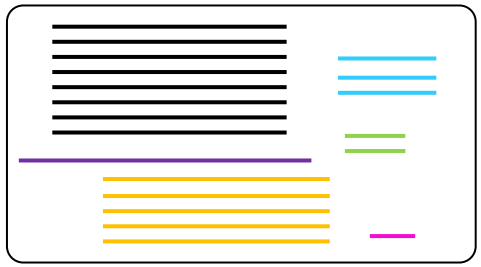


serial	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
0			5							
1	41		58					39		
2	16	17	26		21	16	16	8	24	23
3	235	276	330	297	266	222	213	185	179	199
4	437	456	480	554	507	455	379	300	320	375
5	300	333	381	357	409	371	332	249	452	369
6							27	8	52	25
7	81	82	86	63	83	64	49	27	51	41
8	172	174	84	91	108	105	30	23	41	31
9	11	16		6	11	13	15		22	17
10	173	191	245	255	266	274	359	289	424	384
11	84	91	169	144	96	71	164	114	163	144
12		44	73	64	62	64	60	30	53	43
13	39				37					
14	87	102	85	85	75	64	77	58	101	76
15	48	53	93	60	95	66	55	28	57	45
16	13	15	18	21	22	16	24	27	34	25
17			7				7			
18	21	23	24	28	18	18	23	14	24	27

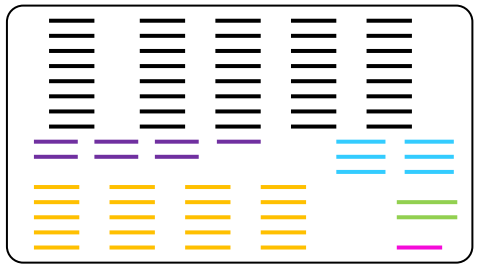
塩基配列データ → 遺伝子発現行列

- 次世代シーケンサー（Illumina社の場合）

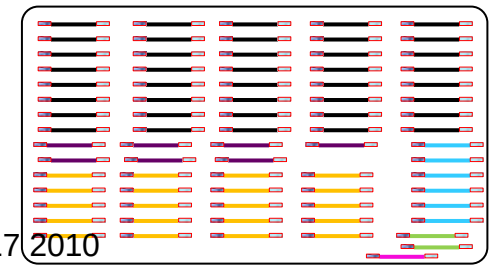
光刺激前（T1）の目のトランスクリプトーム



数百塩基程度
に断片化



二種類のアダプター
配列を両末端に付加

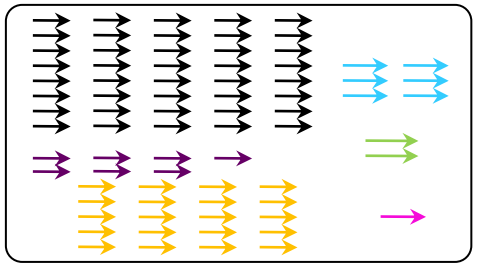


配列決定

・ペアエンド法
断片配列の両末端が数百塩基以内の対の二種類の配列が得られる

・シングルエンド法

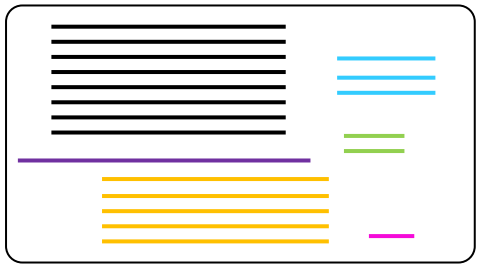
シングルエンド法
の場合



塩基配列データ → 遺伝子発現行列

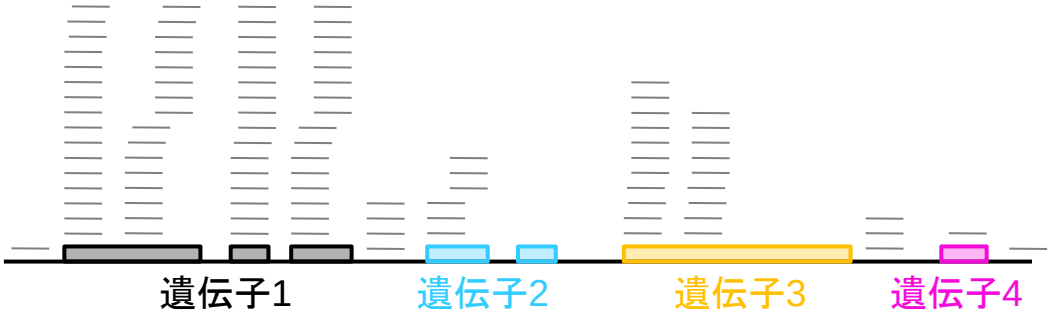
- 次世代シーケンサー（Illumina社の場合）

光刺激前（T1）の目のトランスクリプトーム



シングルエンド解析

ゲノム配列にマッピング



生のリード数をカウント

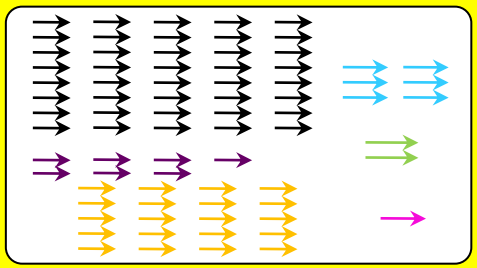
	T1
遺伝子1	40
遺伝子2	6
遺伝子3	20
遺伝子4	1
遺伝子5	…
…	…

正規化

T1
8
3
5
1
…
…

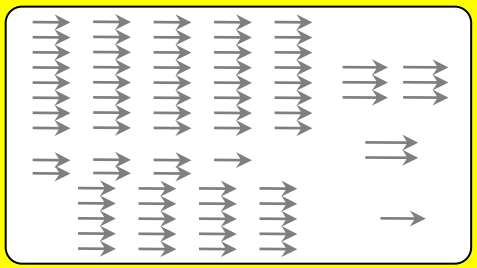
ーイメージー

50-125塩基程度からなる配列が沢山ある



ー実際ー

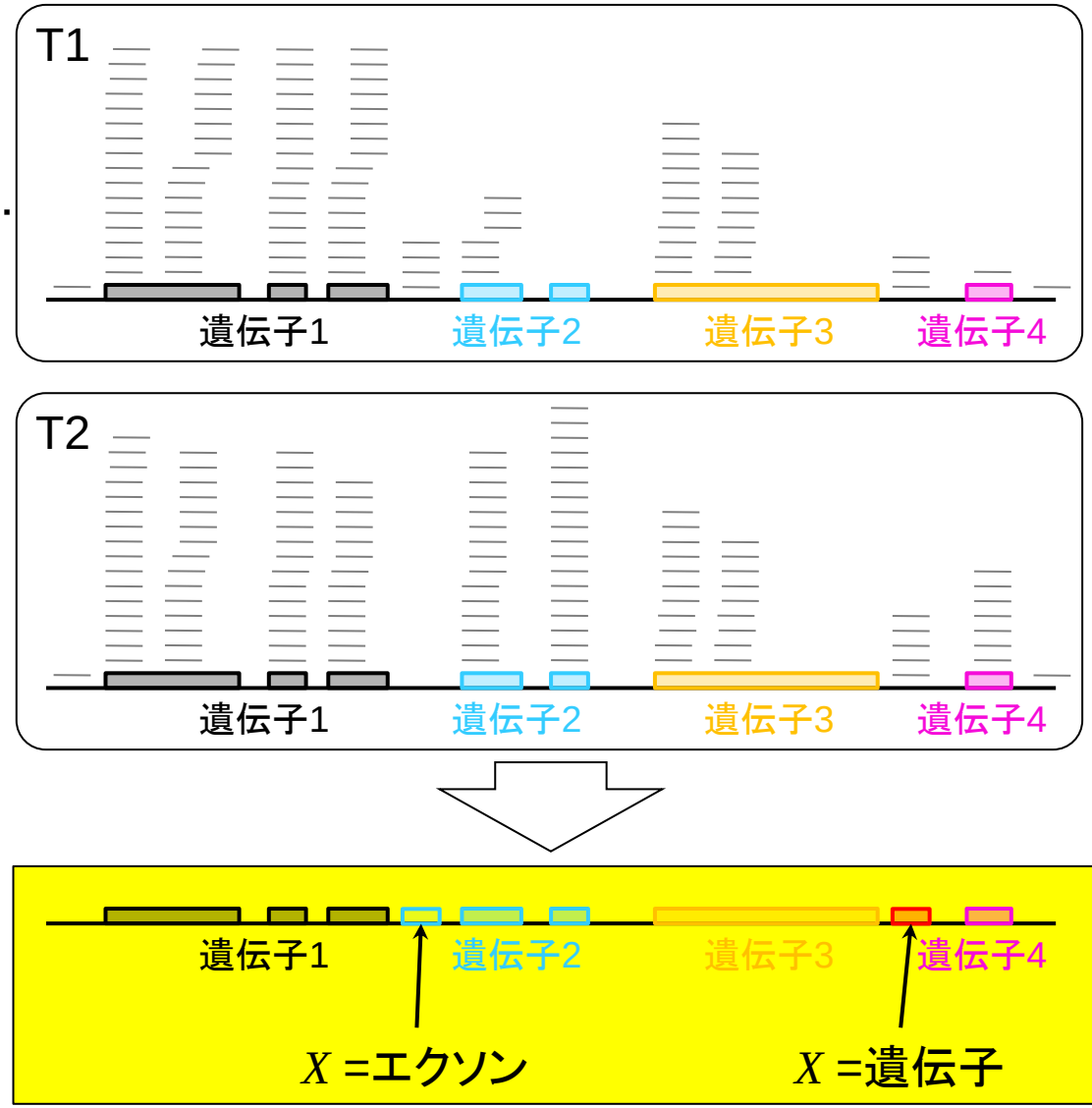
数百万個の配列があり、どの遺伝子に対応するか不明



（短い）配列を読んだものという意味で（ショート）リードなどと呼ばれる

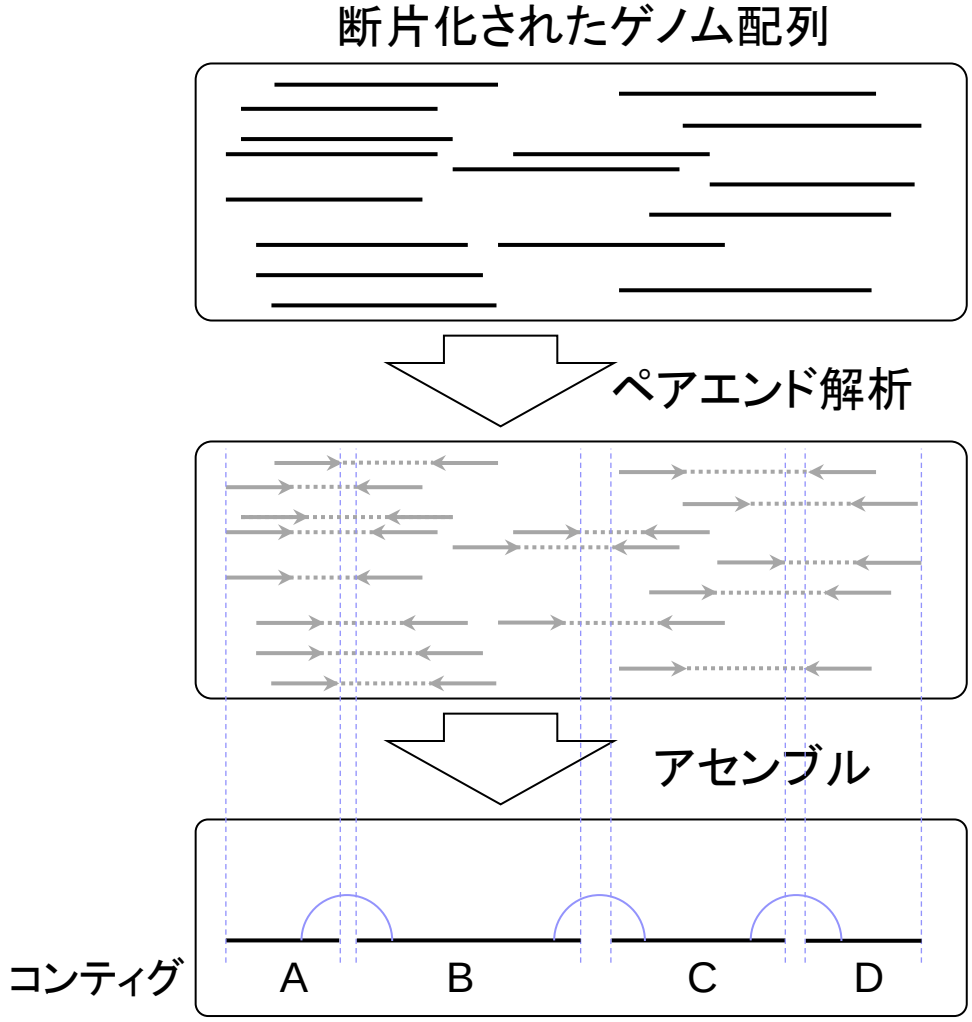
次世代シーケンサー応用例

- 新規Xの同定
  - X =スプライスバリアント, 遺伝子, …



次世代シーケンサー応用例(理論上は?!)

- ゲノム配列そのものを決定(de novo assembly)



コンティグ:連結された断片配列

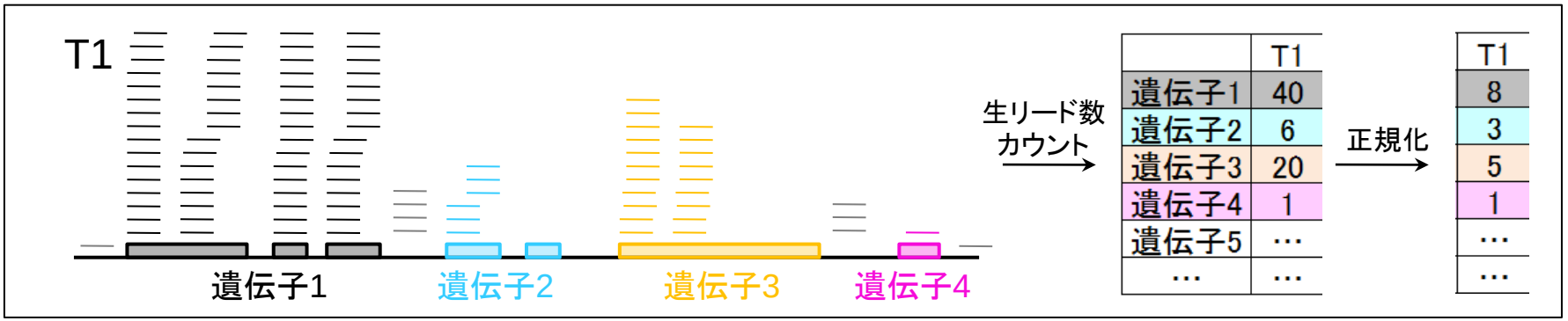
ペアエンド解析のほうがシングルエンド解析よりもコンティグ間の関係(scaffoldまたはsupercontigという)を同定しやすい。
例:「A-D-B-C」ではなく「A-B-C-D」という関係

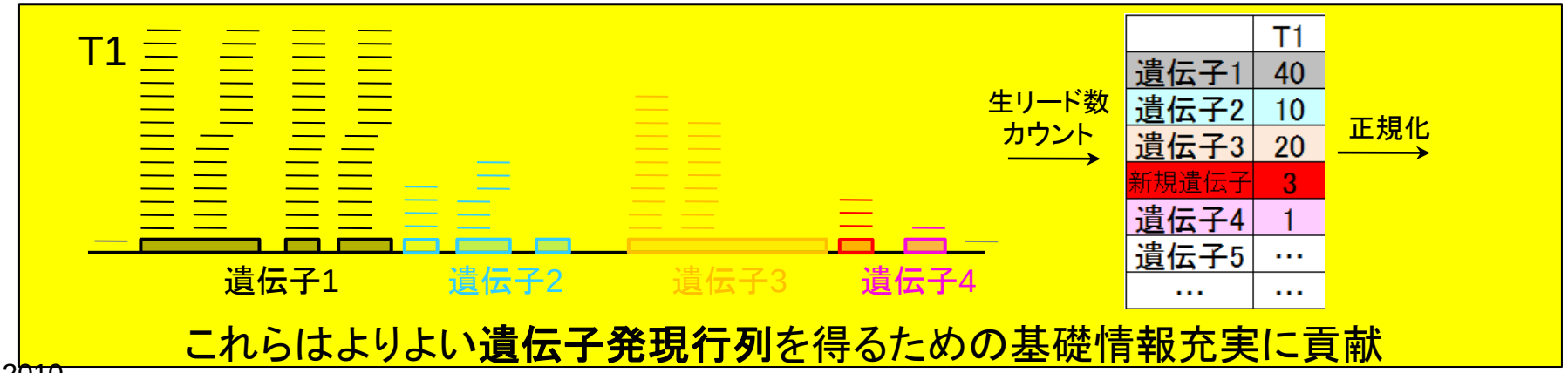
一般論:
アセンブルされたコンティグの正確さは、読んだ配列長が長いほど向上する

次世代シーケンサの特徴:
沢山読めるが配列長(矢印の長さに相当)が短いので、コンティグ形成時の配列間のオーバーラップが短い…

配列決定に基づくトランスクリプトーム解析

- 「ゲノム配列そのものを決定」
- 「新規Xの同定」





配列決定に基づくトランスクリプトーム解析

- ゲノムへのマッピングまではほぼ確立
  - マップされる側のリファレンスゲノム配列をゲット
  - 次世代シーケンサーから得られたマップする側の塩基配列をゲット
  - 上記二つのファイルを入力としてマッピングプログラムを実行

    →SAM/BAM形式の出力ファイルを得る

(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2010/7/20)

What's new?
・(Rで)マイクロアレイデータ解析中の解析法もここで使う予定(あるいはそちらにリンクさせる予定)です(2010/6/3)
・思いつくままにつらつらと書いています。ある程度たまってきたら、項目名を含め大幅にリニューアルする予定です(2010/5/27)
・どこまでできるかわかりませんが次世代シーケンサーのデータ解析(特にトランスクリプトームの方)用のRパッケージの使い方を紹介していきます(2010/5/21)

- はじめに (last modified 2010/7/6)
- Rのインストールと起動 (last modified 2010/7/16) NEW
- イントロダクション | NGS | 各種覚書 (last modified 2010/7/14) NEW
- イントロダクション | NGS | 様々なプラットフォーム (last modified 2010/7/7)
- イントロダクション | NGS | リファレンス配列取得(マップされる側) (last modified 2010/7/9)
- イントロダクション | NGS | アノテーション情報取得(refFlatファイル) (last modified 2010/7/9)
- イントロダクション | 一般 | 配列取得 (last modified 2010/7/7)
- イントロダクション | NGS | Query用配列取得(マップする側) (last modified 2010/7/13) NEW
- イントロダクション | NGS | マッピングプログラム (last modified 2010/7/15) NEW
- イントロダクション | NGS | マッピングプログラムの入力形式について (last modified 2010/7/13) NEW
- イントロダクション | NGS | マッピングプログラムの出力形式について (last modified 2010/7/13) NEW

ゲノムにマップ

- 実データ（ヒトの場合）

マップされる側のリファレンスゲノム配列

```
ファイル(F)  編集(E)  設定(S)  コントロール(O)  ウィンドウ
>chr1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
"reference" 61913917L, 3157608038C
```

・1-22番染色体＋X＋Y
・約6200万行のファイル
・約3GBのサイズ

chr1
chr2
…

マップする側の塩基配列(FASTQ形式)

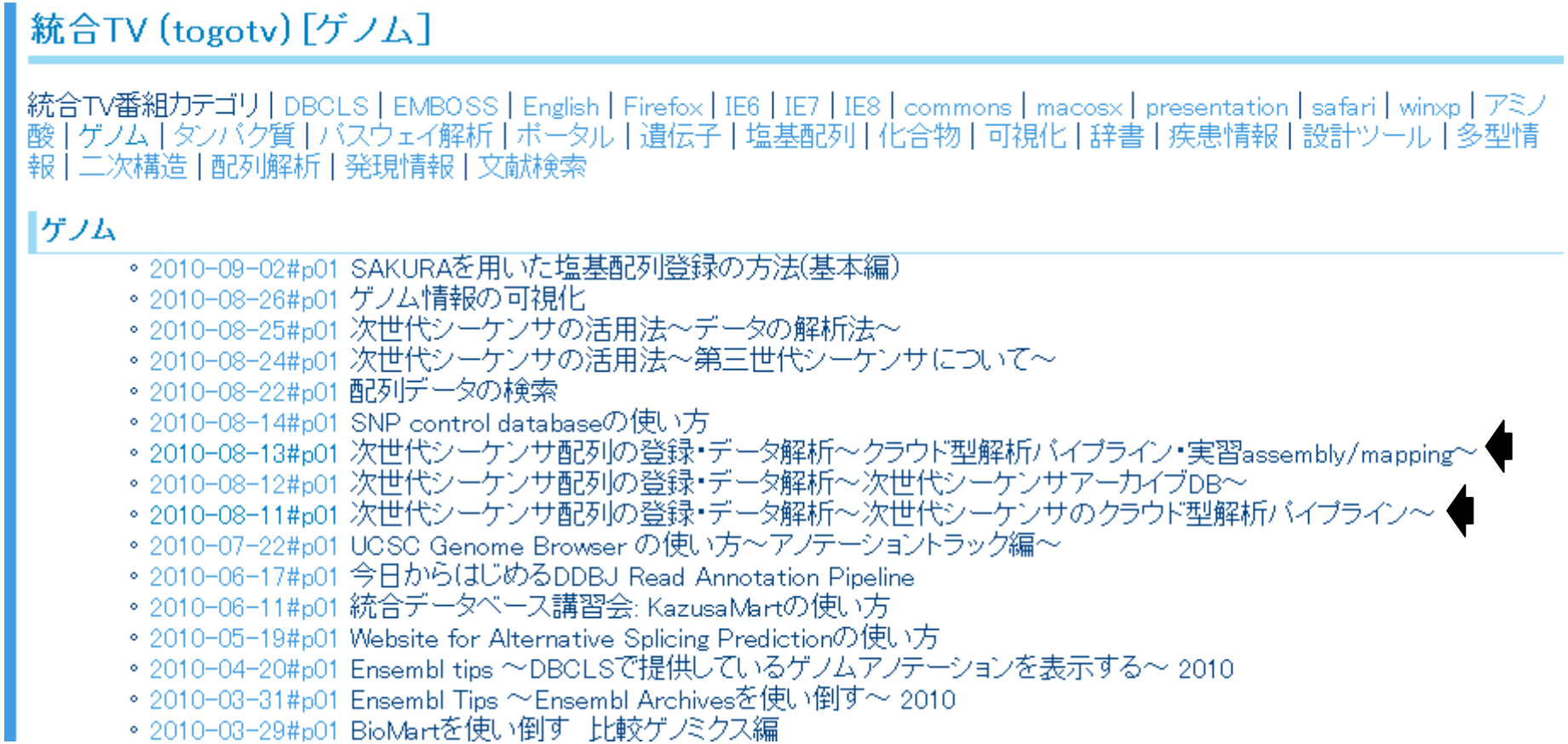
```
ファイル(F)  編集(E)  設定(S)  コントロール(O)  ウィンド
@SRR035678.52 HWI-E4_9_30WAF:1:1:0:1195 length=35
CNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+SRR035678.52 HWI-E4_9_30WAF:1:1:0:1195 length=35
/!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@SRR035678.53 HWI-E4_9_30WAF:1:1:1:196 length=35
ANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+SRR035678.53 HWI-E4_9_30WAF:1:1:1:196 length=35
I!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@SRR035678.54 HWI-E4_9_30WAF:1:1:1:379 length=35
CNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+SRR035678.54 HWI-E4_9_30WAF:1:1:1:379 length=35
I!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@SRR035678.55 HWI-E4_9_30WAF:1:1:1:40 length=35
TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+SRR035678.55 HWI-E4_9_30WAF:1:1:1:40 length=35
I!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@SRR035678.56 HWI-E4_9_30WAF:1:1:1:114 length=35
TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+SRR035678.56 HWI-E4_9_30WAF:1:1:1:114 length=35
I!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@SRR035678.57 HWI-E4_9_30WAF:1:1:1:153 length=35
TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+SRR035678.57 HWI-E4_9_30WAF:1:1:1:153 length=35
I!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@SRR035678.58 HWI-E4_9_30WAF:1:1:1:1122 length=35
TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
"SRR035678.fastq" 46808924L, 2195550366C
```

・約47万行
・1 配列(1 read)を4行で表現
・1配列35塩基長（この場合）

各readが染色体上のどこに一致するかという座標情報を出力するのがマッピングプログラム

DDBJの利用を推奨

- 今後得られるデータ量（1サンプルあたり数百GB）を考慮すると、DDBJ Read Annotation Pipelineを利用するのが手っ取り早いと思います（or 利用する以外の選択肢がない状況になりつつあります）



統合TV (togotv) [ゲノム]

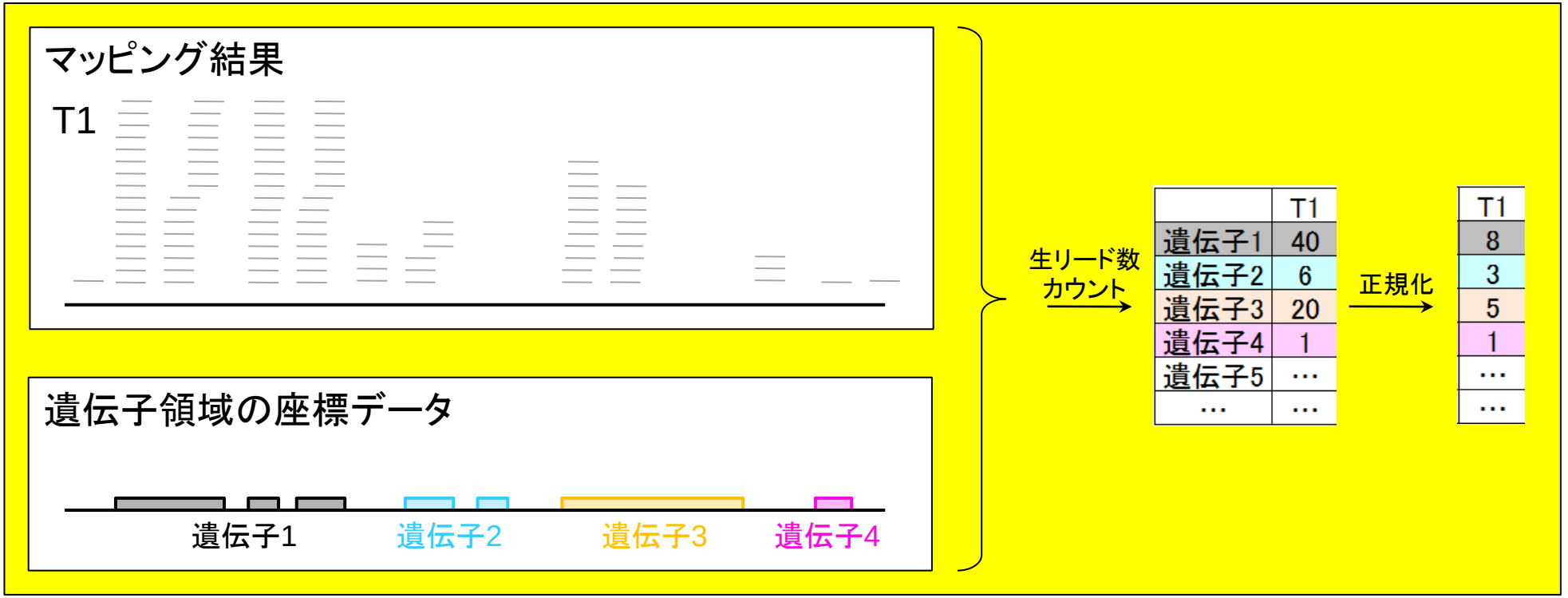
統合TV番組カテゴリ | DBCLS | EMBOSS | English | Firefox | IE6 | IE7 | IE8 | commons | macosx | presentation | safari | winxp | アミノ酸 | ゲノム | タンパク質 | パスウェイ解析 | ポータル | 遺伝子 | 塩基配列 | 化合物 | 可視化 | 辞書 | 疾患情報 | 設計ツール | 多型情報 | 二次構造 | 配列解析 | 発現情報 | 文献検索

ゲノム

- 2010-09-02#p01 SAKURAを用いた塩基配列登録の方法(基本編)
- 2010-08-26#p01 ゲノム情報の可視化
- 2010-08-25#p01 次世代シーケンサの活用法～データの解析法～
- 2010-08-24#p01 次世代シーケンサの活用法～第三世代シーケンサについて～
- 2010-08-22#p01 配列データの検索
- 2010-08-14#p01 SNP control databaseの使い方
- 2010-08-13#p01 次世代シーケンサ配列の登録・データ解析～クラウド型解析パイプライン・実習assembly/mapping～
- 2010-08-12#p01 次世代シーケンサ配列の登録・データ解析～次世代シーケンサアーカイブDB～
- 2010-08-11#p01 次世代シーケンサ配列の登録・データ解析～次世代シーケンサのクラウド型解析パイプライン～
- 2010-07-22#p01 UCSC Genome Browser の使い方～アノテーショントラック編～
- 2010-06-17#p01 今日からはじめるDDBJ Read Annotation Pipeline
- 2010-06-11#p01 統合データベース講習会: KazusaMartの使い方
- 2010-05-19#p01 Website for Alternative Splicing Predictionの使い方
- 2010-04-20#p01 Ensembl tips ～DBCLSで提供しているゲノムアノテーションを表示する～ 2010
- 2010-03-31#p01 Ensembl Tips ～Ensembl Archivesを使い倒す～ 2010
- 2010-03-29#p01 BioMartを使い倒す　比較ゲノミクス編

塩基配列データ → 遺伝子発現行列

- 遺伝子領域の座標データがないと遺伝子発現行列は作れない



	T1
遺伝子1	40
遺伝子2	6
遺伝子3	20
遺伝子4	1
遺伝子5	…
…	…

T1
8
3
5
1
…
…

(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2010/7/20)

What's new?

- イントロダクション | NGS | 各種覚書 (last modified 2010/7/14) NEW
- イントロダクション | NGS | 様々なプラットフォーム (last modified 2010/7/7)
- イントロダクション | NGS | リファレンス配列取得(マップされる側) (last modified 2010/7/9)
- イントロダクション | NGS | アノテーション情報取得(refFlatファイル) (last modified 2010/7/9)
- イントロダクション | 一般 | 配列取得 (last modified 2010/7/7)

塩基配列データ → 遺伝子発現行列

- 遺伝子領域の座標データファイル（例：refFlat形式）

	A	B	C	D	E	F	G	H	I	J	K
1	SNAR-G2	NR_024244	chr19	-	49534925	49535044	49535044	49535044	1	49534925,	49535044,
2	SNAR-D	NR_024243	chr19	-	50643458	50643577	50643577	50643577	1	50643458,	50643577,
3	SNORD113-5	NR_003233	chr14	+	101404523	101404600	101404600	101404600	1	101404523,	101404600,
4	UBL5	NM_024292	chr19	+	9938567	9940797	9939013	9940684	5	9938567,9939002,9939267,9939512,9940640,	9938740,9939069

A: 遺伝子シンボル
B: 遺伝子名
C: 染色体番号
D: 鎖の向き(+鎖 or –鎖)
E：転写開始位置
F：転写終結位置
G：コーディング領域の開始位置
H：コーディング領域の終結位置
I：エクソンの数
J：エクソンの開始位置
K：エクソンの終結位置

座標データファイルも無料で公開されている

塩基配列データ → 遺伝子発現行列

(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2010/7/20)

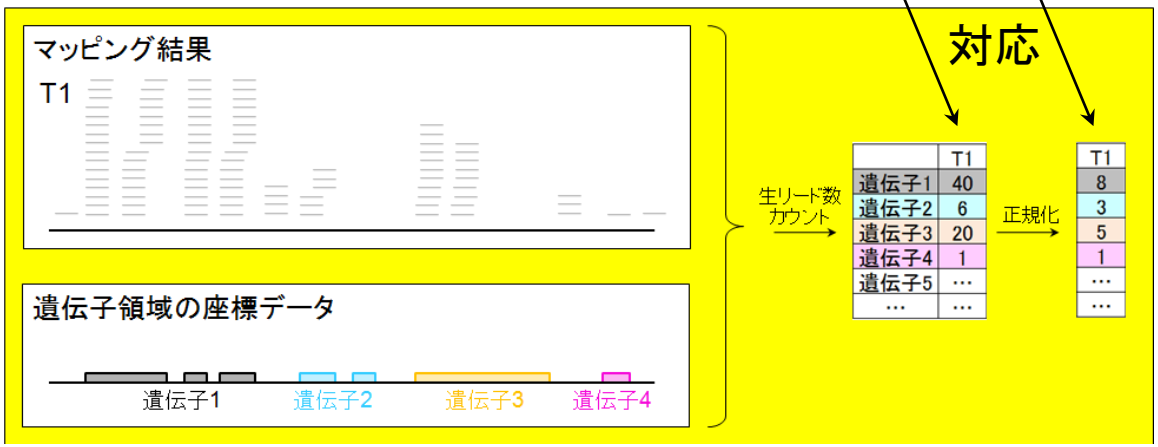
What

重複なし) (last modified 2010/6/8)

- 前処理 | ゲノムへのマッピング結果から既知遺伝子の発現レベル(RPKM)への変換 (last modified 2010/9/14) NEW
- 前処理 | サンプル間比較を行うための正規化について(RPM, RPKM, ...) (last modified 2010/7/20)

の結果ファイル

geneName	raw counts	RPKM	all reads	gene length
A1BG	744	82.9	5087097	1764
A1CF	159	13.7	5087097	2278
A2BP1	1	0.0	5087097	5415
A2LD1	4	0.6	5087097	1226
A2M	2373	100.3	5087097	4653



このサンプルを次世代シーケンサーにかけると**5087097** reads (重複を含む塩基配列数)からなるデータが得られており、そのうち**744** readsが**A1BG**という遺伝子上にマップされていて、この遺伝子の正規化後の発現レベルは**82.9** RPKMですよ。

データの正規化

- RPM正規化（マイクロアレイなどと同じところ）
  - Reads per million mapped readsの略
  - サンプルごとに読まれた総リード（塩基配列）数が異なる。
    →各遺伝子のマップされたリード数を「総read数が100万（one million）だった場合」に補正

「生read数 : 総read数 = x : 1,000,000 」
A1BGの場合は「744 : 5,087,097 = x : 1,000,000」

$$x = \text{生read数} \times \frac{1000000}{\text{総read数}} = 744 \times \frac{1000000}{5087097} = 146.3$$

geneName	raw counts	RPKM	all reads	gene length	RPM
A1BG	744	82.9	5087097	1764	146.3
A1CF	159	13.7	5087097	2278	31.3
A2BP1	1	0.0	5087097	5415	0.2
A2LD1	4	0.6	5087097	1226	0.8

- RPKM正規化（次世代シーケンサ特有）
  - Reads per kilobase of exon per million mapped readsの略
  - 遺伝子の配列長が長いほど配列決定(sequence)される確率が上昇
    →各遺伝子の配列長を「1000塩基（one kilobase）だった場合」に補正

$$\text{生read数} \times \frac{1000000}{\text{総read数}} \times \frac{1000}{\text{配列長}} = 744 \times \frac{1000000}{5087097} \times \frac{1000}{1764} = 82.9$$

遺伝子発現行列 → 様々な解析が可能

- RPKM正規化後の遺伝子発現行列（ファイル名：data.txt）

14サンプル
（A: 7サンプル、B：7サンプル）

21,717遺伝子

symbol	A1	A2	A3	A4	A5	A6	A7	B1	B2	B3	B4	B5	B6	B7
A1BG	7.2	7.7	7.6	8.0	7.6	7.1	7.1	3.5	4.4	3.6	3.9	4.2	4.2	4.0
A1CF	3.1	3.4	2.9	3.1	3.8	2.5	3.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0
A2BP1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	26.3	23.7	23.1	23.7	25.4	24.3	23.5
A2LD1	1.6	2.1	2.7	2.5	1.8	2.4	1.5	1.2	1.4	1.4	1.2	0.9	1.9	1.0
A2M	93.8	94.9	91.3	91.5	91.8	94.3	93.7	23.1	23.5	22.6	23.6	22.4	23.3	24.6
A2ML1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.3	0.4	0.5	0.2	0.2	0.4	0.3
A4GALT	7.2	7.6	7.2	5.1	7.1	7.9	6.1	3.6	3.5	4.5	4.0	2.9	3.8	3.6
A4GNT	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0
AAA1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AAAS	33.0	33.5	33.8	34.6	33.0	34.6	33.7	12.4	11.2	11.8	12.0	12.3	13.3	13.4
AACS	12.4	13.1	13.4	12.0	12.9	12.6	12.5	14.9	13.6	11.9	13.4	14.0	14.1	13.2
AACSL	0.3	0.4	0.9	0.8	0.8	0.8	0.6	0.0	0.0	0.1	0.0	0.1	0.2	0.1
AADAC	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AADACL2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AADACL3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AADACL4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AADAT	2.2	1.8	1.9	2.0	1.2	2.4	2.0	1.4	1.3	1.7	0.9	2.2	1.2	1.7
AAGAB	14.0	15.8	13.7	13.9	15.6	15.6	15.3	6.6	6.4	6.3	6.1	7.1	6.7	6.3
...														

参考

(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2010/7/20)

What's new?
・(Rで)マイクロアレイデータ解析中の解析法もここで使う予定(あるいはそちらにリンクさせる予定)です(2010/6/3)
・思いつくままにつらつらと書いています。ある程度たまってきたら、項目名を含め大幅にリニューアルする予定です(2010/5/27)
・どこまでできるかわかりませんが次世代シーケンサーのデータ解析(特にトランスクリプトームの方)用のRパッケージの使い方を紹介していきます(2010/5/21)

- はじめに (last modified 2010/7/6)
- Rのインストールと起動(last modified 2010/7/16) NEW
- イントロダクション | NGS | 各種覚書(last modified 2010/7/14) NEW
- イントロダクション | NGS | 様々なプラットフォーム(last modified 2010/7/7)
- イントロダクション | NGS | リファレンス配列取得(マップされる側)(last modified 2010/7/9)
- イントロ
- イントロ
- イントロ
- イントロ
- イントロ

(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2010/09/01)

What's new?
・GSA (Efron_2007)の中身をちゃんと埋め始めましたが、まだ最後までは辿りつけてません(2010/8/30)
・Hook (Binder_2008)を追加しました(2010/8/10)
・Agilent two-color processing用のRパッケージを偶然発見したので(項目のみですが...)追加しました(2010/7/14)
・作図 | ROC曲線(ROC curve)を追加しました(2010/4/20)
・このページとは直接関係ありませんが、(Rで)塩基配列解析というページで主に次世代シーケンサーデータ解析を意識したページを作成しつつありますので、そっち方面の解析をRでやりたい方はそちらをご覧ください(2010/5/27)
・Rのインストールと起動のところに64 bitパソコンの場合を追加しました(2010/5/21)
・作図 | ROC曲線(ROC curve)を追加しました(2010/4/20)
・Linksのところにこのページの解析結果?!を可視化させるためのプラットフォーム情報などを追加しました(2010/4/20)
・ヒートマップのところのリンク切れなどを修正しました(2010/4/9)
・解析 | クラスタリング | 階層的 | hclustのところに、自動的に図をpng形式で出力するやり方を掲載しました(2010/1/29)
・アレイCGH(DNAコピー数)解析のやり方を掲載すべく善処しております。(2009/12/14)
・2009年11月20, 24日13:00-マイクロアレイデータ解析講習会を開催しました。
・発現変動遺伝子でないものの割合をざっと調べるための手段などを掲載しました。(2009/11/6)
・入力データ中に「'」があるとちゃんと読み込めなかった(例:「2'-PDE」というGene symbolの行で読み込みがストップしてしまう)のでread.table関数での読み込み時に「quote="'"」というオプションを追加しました。(2009/10/16)

解析例1（サンプル間クラスタリング）

(Rで)マイクロアレイデータ解析 by 門田幸二 (last modified 2010/09/01)

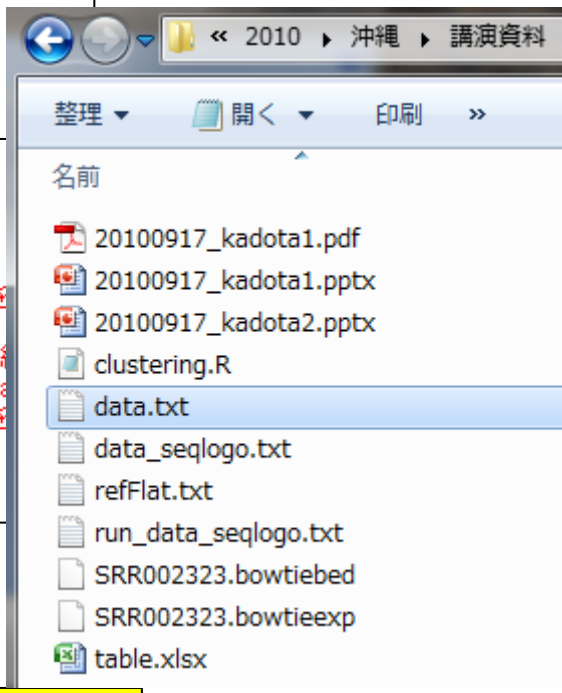
What's new?
- GSA (Efron_2007)の中身をちゃんと埋め始めましたが、まだ最後までは辿りつけてません(2010/8/30)
- Hook (Binder_2008)を追加しました(2010/8/10)

- 解析 | クラスタリング | 階層的 | について(last modified 2009/8/12)
- 解析 | クラスタリング | 階層的 | pvclust (Suzuki_2006) (last modified 2010/8/5) NEW
- 解析 | クラスタリング | 階層的 | hclust (last modified 2010/1/29)
- 解析 | クラスタリング | 階層的 | hclust後の詳細な解析 (last modified 2009/8/7)
- 解析 | クラスタリング | 階層的 | 最適なクラスター数を見積る (last modified 2009/9/9)
- 解析 | クラスタリング | 非階層的 | K-means
- 解析 | クラスタリング | 非階層的 | 自己組織化マップ(SOM)

2. サンプル間クラスタリングの場合(類似度:「1-相関係数」、方法:平均連結法(average)):

・R Graphics画面上に表示したい場合：

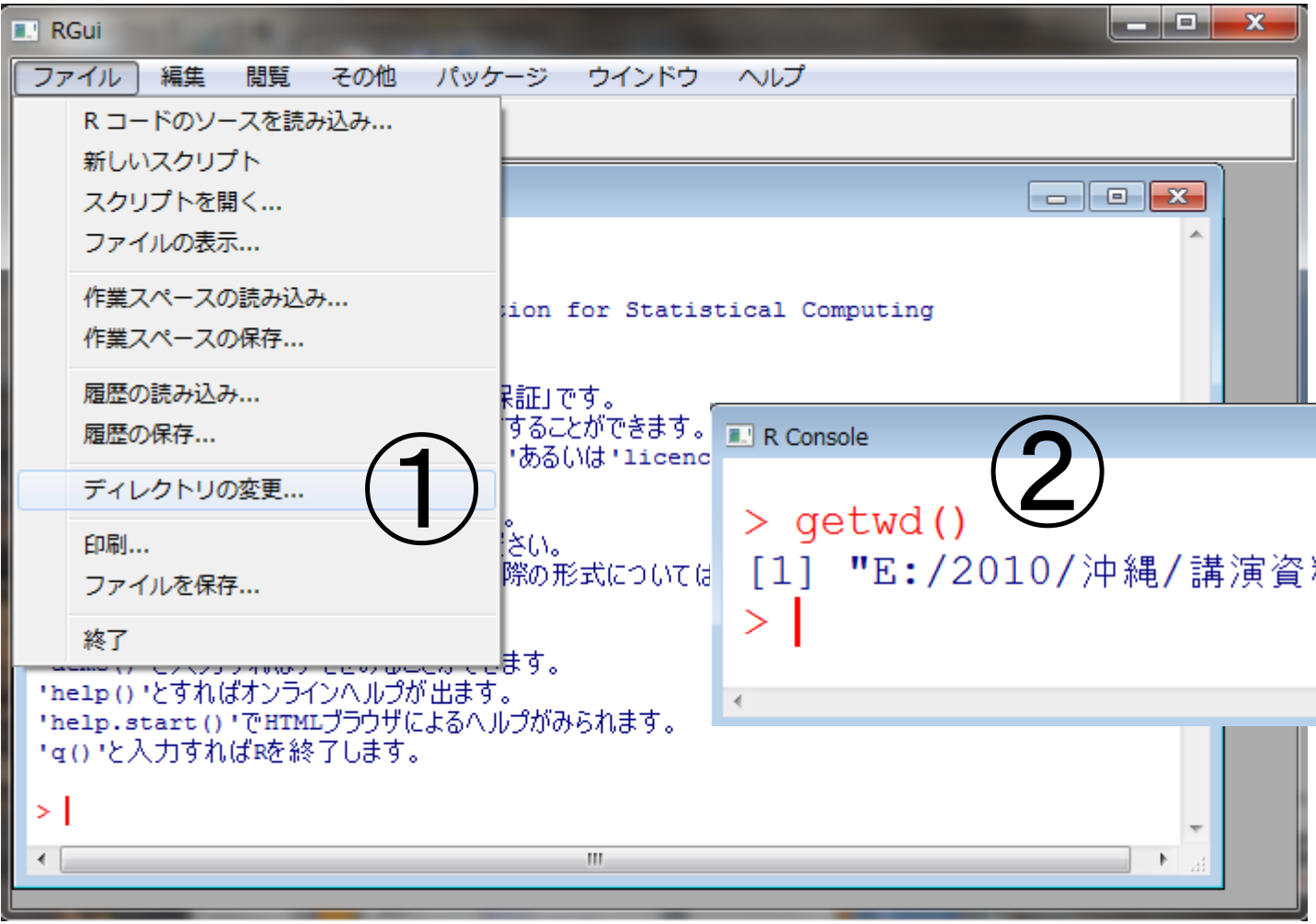
```
------    ここから    ------
in_f <- "sample3.txt"                                          #入力ファイル名(発現データファイル)を
param2 <- "average"                                            #方法(method)を指定
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#発現データを読み込んでdataに格
data.dist <- as.dist(1 - cor(data))                            #サンプル間の距離を計算し、結果をdata
out <- hclust(data.dist, method=param2)                        #階層的クラスタリングを実行し、結果を
plot(out)                                                      #樹形図（デンドログラム）の表示

------    ここまで    ------
```



解析したいのは「… – 2010 – 沖縄 – 講演資料」フォルダ中の「data.txt」ファイル

解析例1（サンプル間クラスタリング）

①Rを起動し、「ファイル」ー「ディレクトリの変更」で解析したいファイル(data.txt)を置いてあるディレクトリに移動。②念のため確認



解析例1（サンプル間クラスタリング）

③入力ファイル名の部分を変更したものを用意し、④R Console上でコピペ

③

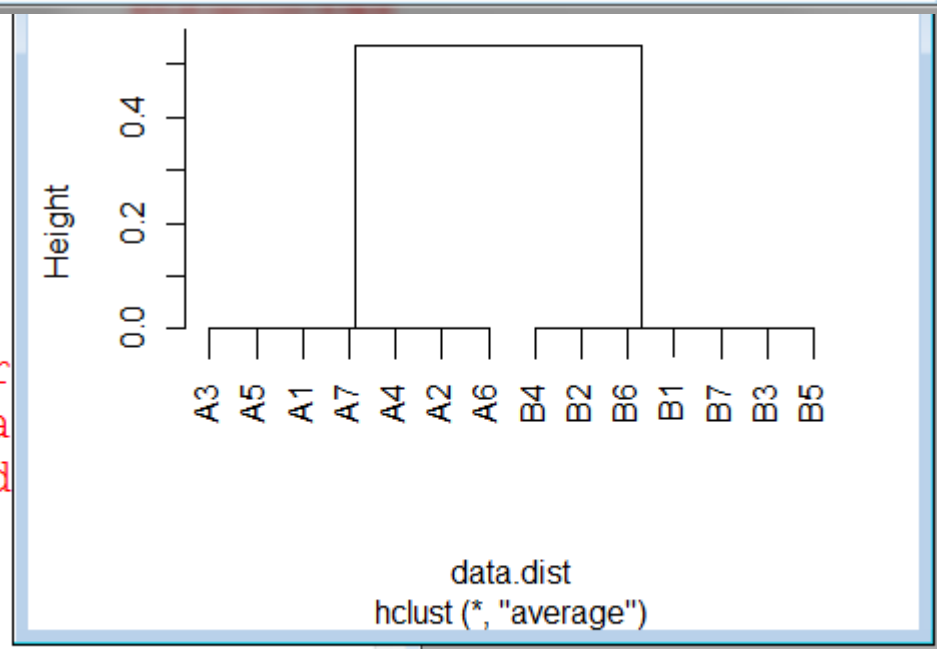
clustering.R - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
in_f <- "data.txt"                                          #入力ファイル名(発現データファイル)を指定
param2 <- "average"                                         #方法(method)を指定
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#発現データを読み込んでdataに格納。
data.dist <- as.dist(1 - cor(data))                         #サンプル間の距離を計算し、結果をdata.distに格納
out <- hclust(data.dist, method=param2)                     #階層的クラスタリングを実行し、結果をoutに格納
plot(out)                                                   #樹形図（デンドログラム）の表示
```

④

```
'q()'と入力すればRを終了します。

> getwd()
[1] "E:/2010/沖縄/講演資料"
> in_f <- "data.txt"
> param2 <- "average"
> data <- read.table(in_f, header
> data.dist <- as.dist(1 - cor(da
> out <- hclust(data.dist, method
> plot(out)
>
```



距離(類似度)の定義

- 遺伝子(or サンプル)xとyの発現パターンの距離D

相関係数 $r_{xy} = \dfrac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ $\quad (-1 \le r_{xy} \le 1)$

i	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
5	x_5	y_5
…	…	…
n	x_n	y_n

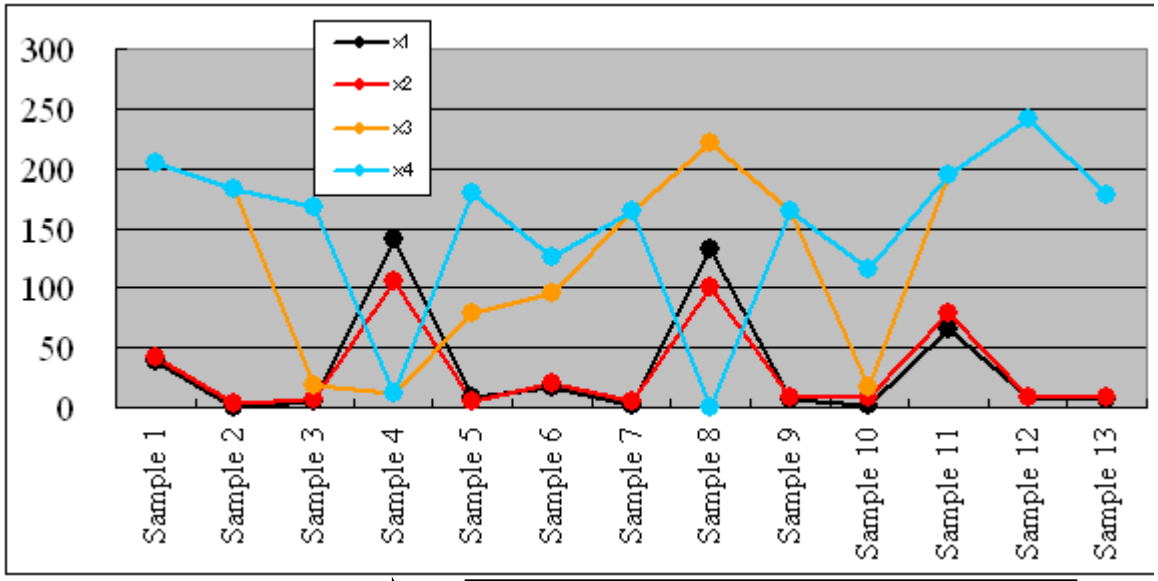
- xとyの発現パターンが酷似 $\rightarrow$ $r \approx 1$
- xとyの発現パターンがばらばら $\rightarrow$ $r \approx 0$
- xとyの発現パターンがほぼ正反対 $\rightarrow$ $r \approx -1$

距離 $D = 1 - r \quad (0 \le D \le 2)$

- $r = 1 \rightarrow D = 1 - 1 = 0$
- $r = 0 \rightarrow D = 1 - 0 = 1$
- $r = -1 \rightarrow D = 1 - (-1) = 2$

階層的クラスタリング

1. 遺伝子間距離を計算



例：4遺伝子の場合

	x^1	x^2	x^3	x^4
Sample 1	38	42	204	204
Sample 2	0	3	182	182
Sample 3	6	6	19	168
Sample 4	141	106	11	11
Sample 5	8	5	79	179
Sample 6	16	20	96	126
Sample 7	2	5	164	164
Sample 8	132	101	222	0
Sample 9	7	9	164	164
Sample 10	2	8	16	116
Sample 11	66	79	195	195
Sample 12	8	9	241	241
Sample 13	6	8	177	177

距離 $D = 1 - r \quad (0 \le D \le 2)$ ⇒ 距離 $D = \frac{1-r}{2} \quad (0 \le D \le 1)$

相関係数 $r_{1,2} = 0.98 \rightarrow$ 距離 $D_{1,2} = \frac{1-0.98}{2} = 0.01$

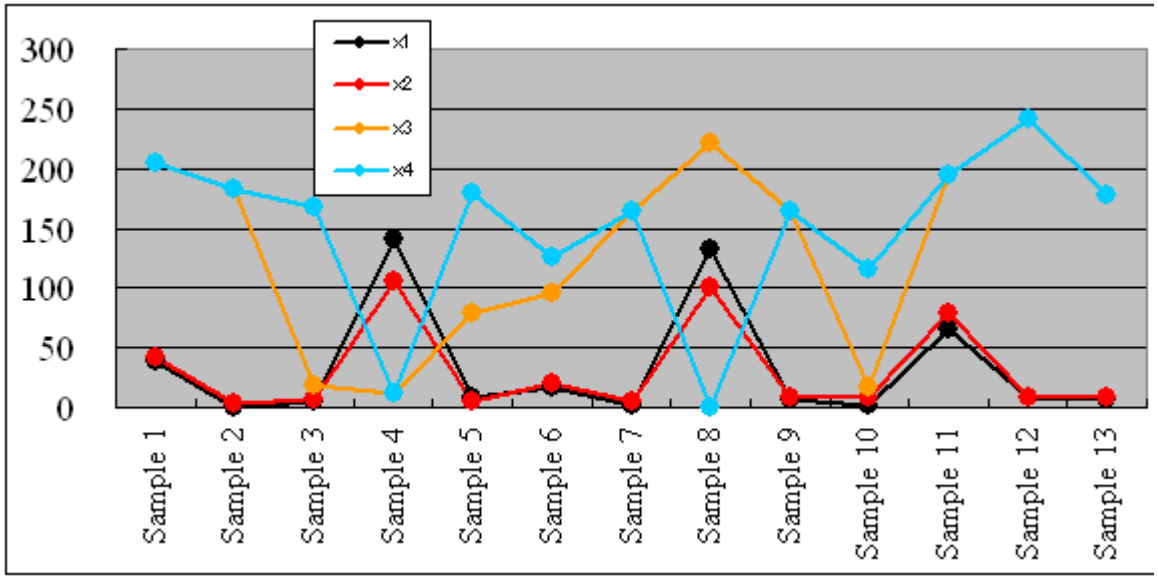
相関係数 $r_{1,3} = -0.01 \rightarrow$ 距離 $D_{1,3} = \frac{1-(-0.01)}{2} = 0.50$

相関係数 $r_{1,4} = -0.78 \rightarrow$ 距離 $D_{1,4} = \frac{1-(-0.78)}{2} = 0.89$

…

階層的クラスタリング

2. 距離行列を作成



距離 $D_{1,2} = \frac{1-0.98}{2} = 0.01$

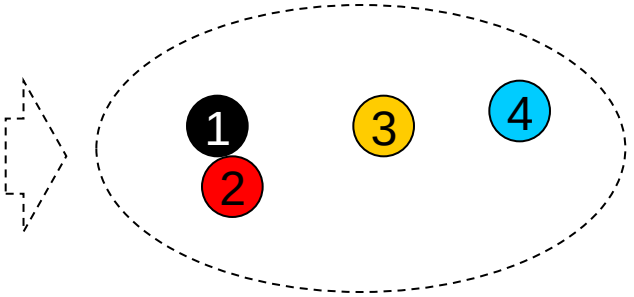
距離 $D_{1,3} = \frac{1-(-0.01)}{2} = 0.50$

距離 $D_{1,4} = \frac{1-(-0.78)}{2} = 0.89$

…

	x^2	x^3	x^4
x^1	0.01	0.50	0.89
x^2		0.47	0.84
x^3			0.32

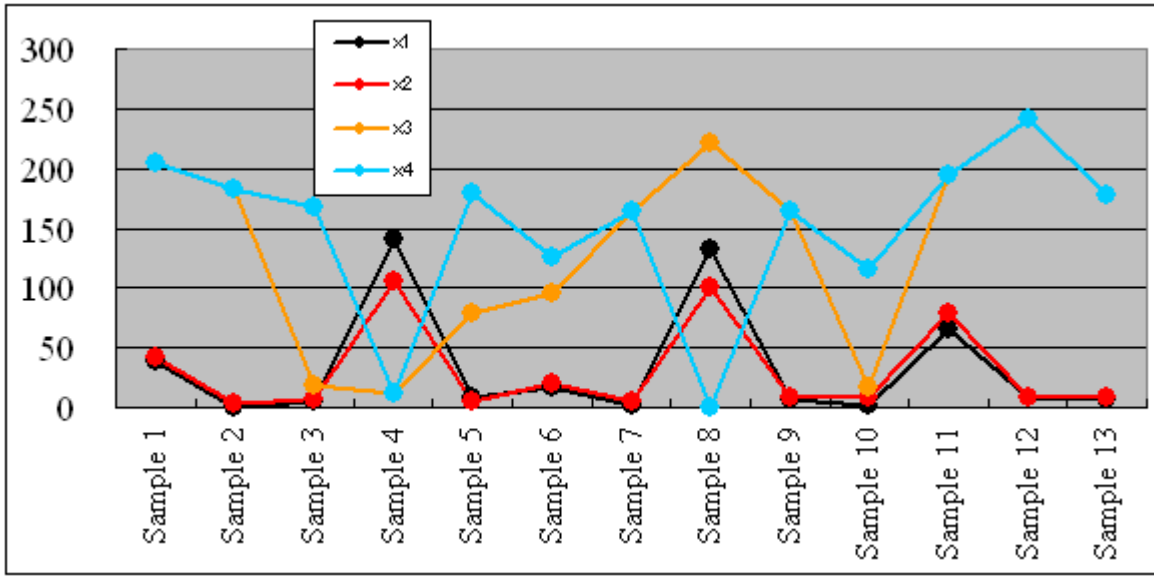
距離行列



イメージ

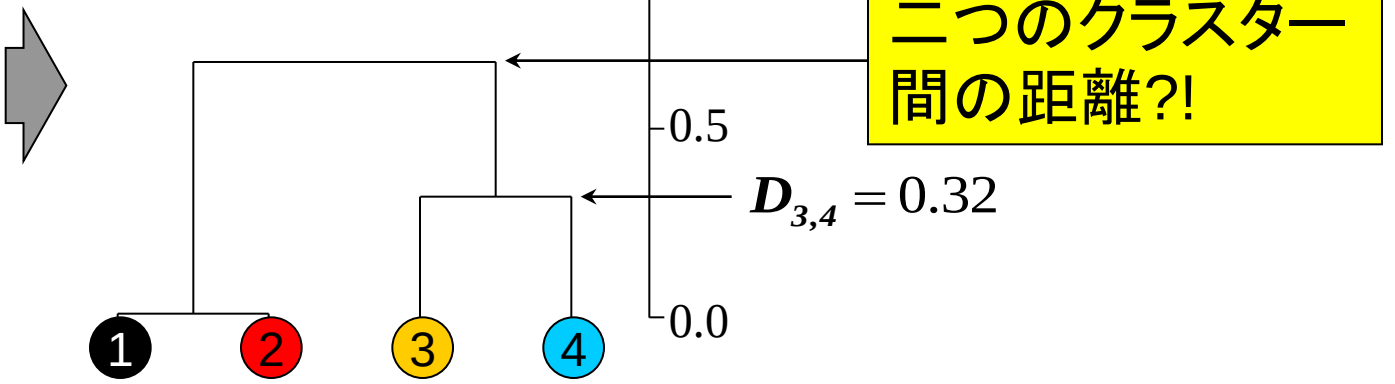
階層的クラスタリング

3. 樹形図を作成



	x^2	x^3	x^4
x^1	0.01	0.50	0.89
x^2		0.47	0.84
x^3			0.32

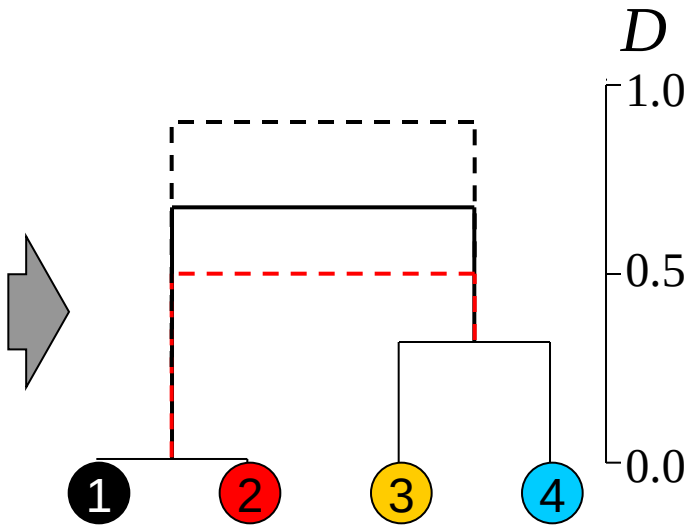
距離行列



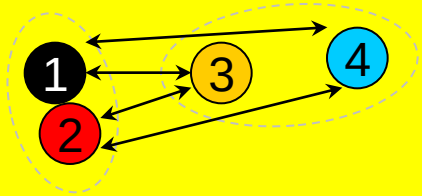
階層的クラスタリング

3. 樹形図を作成

	x^2	x^3	x^4
x^1	0.01	0.50	0.89
x^2		0.47	0.84
x^3			0.32



平均連結法の場合



$$(D_{1,3}+D_{1,4}+D_{2,3}+D_{2,4})/4$$
$$=(0.50+0.89+0.47+0.84)/4$$
$$=0.68$$

単連結法の場合

$$\min(D_{1,3},D_{1,4},D_{2,3},D_{2,4})$$
$$=0.47$$

完全連結法の場合

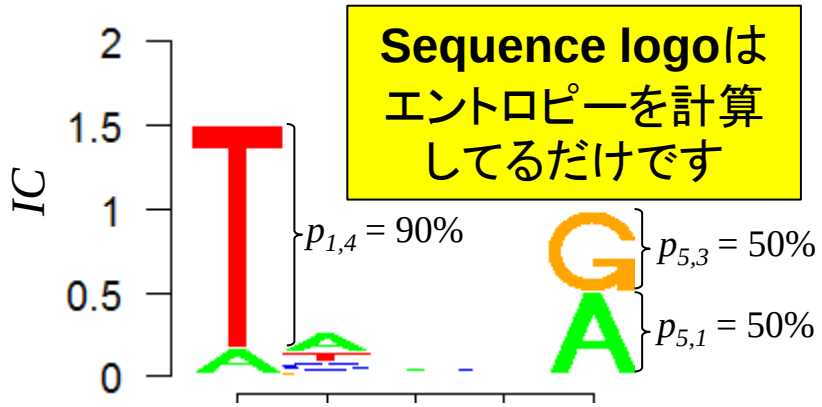
$$\max(D_{1,3},D_{1,4},D_{2,3},D_{2,4})$$
$$=0.89$$

解析例2（Sequence logo）

- position iの情報量 $IC_i = \log_2(N) - H(x_i)$
  $= 2$

N：塩基の種類数= 4
Hの取りうる範囲：$0 \leqq H \leqq \log_2 N$

	position i 1	2	3	4	5	…
配列 1	T	A	C	G	G	…
配列 2	T	A	A	C	G	…
配列 3	T	G	T	A	G	…
配列 4	A	C	T	T	A	…
配列 5	T	T	G	G	A	…
配列 6	T	C	A	A	G	…
配列 7	T	A	C	T	A	…
配列 8	T	T	G	C	A	…
配列 9	T	A	A	C	A	…
配列10	T	A	C	T	G	…



IC	1.53	0.24	0.03	0.03	1.00	…

x_{ij}	1	2	3	4	5	…
Aの数 (j=1)	1	5	3	2	5	…
Cの数 (j=2)	0	2	3	3	0	…
Gの数 (j=3)	0	1	2	2	5	…
Tの数 (j=4)	9	2	2	3	0	…
$\Sigma_j\ x_{ij}$	10	10	10	10	10	

j

p_{ij}	1	2	3	4	5	…
1	0.1	0.5	0.3	0.2	0.5	…
2	0.0	0.2	0.3	0.3	0.0	…
3	0.0	0.1	0.2	0.2	0.5	…
4	0.9	0.2	0.2	0.3	0.0	…
Σ_j	1.0	1.0	1.0	1.0	1.0	

$-p_{ij}\log_2(p_{ij})$	1	2	3	4	5	…
1	0.33	0.50	0.52	0.46	0.50	…
2	0.00	0.46	0.52	0.52	0.00	…
3	0.00	0.33	0.46	0.46	0.50	…
4	0.14	0.46	0.46	0.52	0.00	…
$H = \Sigma_j$	0.47	1.76	1.97	1.97	1.00	

解析例2（Sequence logo）

(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2010/9/15)

What's new?
・(Rで)マイクロアレイデータ解析中の解析法もここで使う予定(あるいはそちらにリンクさせる予定)です(2010/6/3)
・思いつくままにつらつらと書いています。ある程度たまってきたら、項目名を含め大幅にリニューアルする予定です(2010/5/27)

- 解析 | 一般 | アラインメント(ペアワイズ;基本編2) (last modified 2010/6/8)
- 解析 | 一般 | アラインメント(ペアワイズ;応用編) (last modified 2010/6/8)
- 解析 | 一般 | パターンマッチング (last modified 2010/6/30)
- 解析 | 一般 | GC含量 (last modified 2010/7/1)
- 解析 | 一般 | Sequence logos (Schneider_1990) (last modified 2010/9/15) NEW
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | MARS (last modified 2010/7/9)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | MARS (last modified 2010/7/9)

2. 入力ファイルが塩基組成のファイルの場合：

```
------    ここから    ------
in_f <- "data_seqlogo.txt"                    #読み込みたいファイル名を指定してin_fに格納
library(Biostrings)                           #パッケージの読み込み
library(seqLogo)                              #パッケージの読み込み
hoge <- hoge <- read.table(in_f)              #in_fで指定したファイルの読み込み
out <- makePWM(hoge)                          #情報量(information content; ic)を計算している
seqLogo(out)                                  #塩基組成やicの情報を含むoutを入力としてsequence logoを描画。単

------    ここまで    ------
```

解析例2（Sequence logo）

data_seqlogo.txt

0.1	0.5	0.3	0.2	0.5
0	0.2	0.3	0.3	0
0	0.1	0.2	0.2	0.5
0.9	0.2	0.2	0.3	0

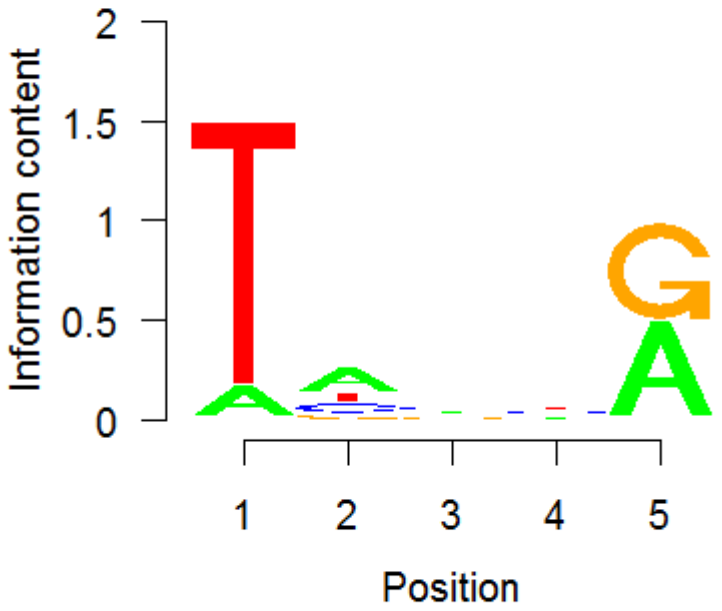
RGui

ファイル　履歴　サイズ変更　ウインドウ

R Console

```
>
> in_f <- "data_seqlogo.txt"                    $
> library(Biostrings)                           $
 要求されたパッケージ IRanges をロード中です

 次のパッケージを付け加えます: 'IRanges'

The following object(s) are masked from 'pack$

    cbind, Map, mapply, order, paste,
    pmax, pmax.int, pmin, pmin.int,
    rbind, rep.int, table

> library(seqLogo)                              $
 要求されたパッケージ grid をロード中です
> hoge <- hoge <- read.table(in_f)              $
> out <- makePWM(hoge)                          $
> seqLogo(out)                                  $
>
>
```

R Graphics: Device 2 (ACTIVE)



参考

使い倒し系チャンネル
統合TV

統合TVは、ライフサイエンス統合データベースセンター(DBCLS)が発信する**生命科学分野の有用なデータベース(DB)やウェブツールの活用法を動画で紹介するウェブサイト**です。くわしくは はじめての方へ をご覧ください。ご意見、ご要望は、お問い合わせ からお願いします。

Google™ +Dashboard widget

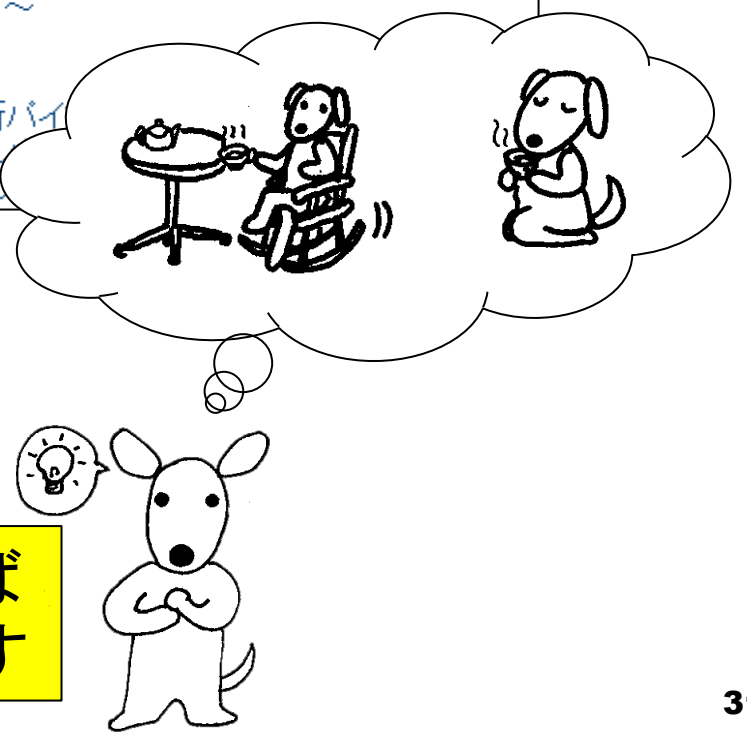
番組検索はこちらから！統合TVまとめサイト 統合TV Curated →

TOGO TV CURATED

番組
質 |
| 文

ゲノム

- 2010-09-02#p01 SAKURAを用いた塩基配列登録の方法(基本編)
- 2010-08-26#p01 ゲノム情報の可視化
- 2010-08-25#p01 次世代シーケンサの活用法～データの解析法～
- 2010-08-24#p01 次世代シーケンサの活用法～第三世代シーケンサについて～
- 2010-08-22#p01 配列データの検索
- 2010-08-14#p01 SNP control databaseの使い方
- 2010-08-13#p01 次世代シーケンサ配列の登録・データ解析～クラウド型解析パイ
- 2010-08-12#p01 次世代シーケンサ配列の登録・データ解析～次世代シーケン
- 2010-08-11#p01 次世代シーケンサ配列の登録・データ解析～次世代シーケン

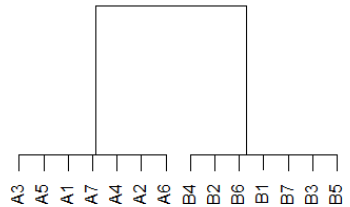


ネット接続環境さえ整っていれば
どこでも情報収集できる時代です

まとめ

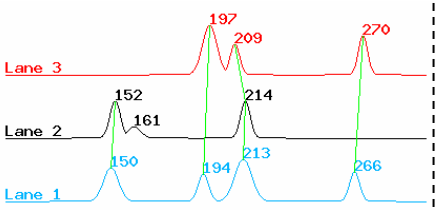
クラスタリング

ー基本編ー



サンプル間

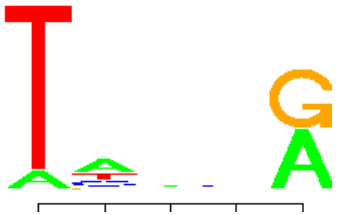
ー応用編ー



ピークマッチング

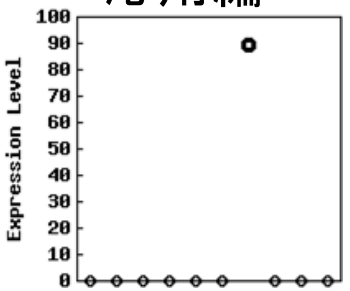
エントロピー

ー基本編ー



sequence logo

ー応用編ー



組織特異的遺伝子

第二部

第一部

次世代シーケンサデータもRのコピペで解析可能
→ 頭脳労働

バイオインフォ要素技術の習得は大事だが、それだけでも様々な種類の実験データに対応可能

10:00-19:00（完全週休二日）の研究生活です

謝辞

東京大学　大学院農学生命科学研究科

清水謙多郎　教授

グラント

- 若手研究(B)（H21年度- ）:「マイクロアレイ解析の再現性・感度・特異度を飛躍的に向上させるデータ解析手法の開発」(代表)
- 新学術領域研究(研究領域提案型)(H22年度-):「非モデル生物におけるゲノム解析法の確立」(分担)